Microbial pathogen genomes – new strategies for identifying therapeutics and vaccine targets

Douglas R. Smith

Advances in high-throughput DNA-sequencing techniques have given us the unprecedented ability to rapidly determine the nucleotide sequences of entire bacterial genomes. The application of these methods to the genomes of microbial pathogens, combined with efficient analytical tools and genome-scale approaches for studying gene expression, is revolutionizing our approach to the selection of targets for drug screening and vaccine development. This is bringing new life to this important, but long-neglected, field of research.

The decision, several years ago, by the US Department of Energy, the National Institutes of Health (NIH) and several international funding agencies to embark upon programs to map and sequence the human genome has led to a number of important technological advances that are beginning to have an impact in other areas of biology. Among these advances are the development of automated methods for the generation of large amounts of raw DNA-sequencing information, computer software for rapidly processing and analyzing primary sequence data, and techniques for the rapid assembly of shotgun sequencing reads, even from entire bacterial genomes. Efficient algorithms for similarity searching allow the rapid identification of protein-encoding sequences that are homologous to other genes, the sequences of which are held in public and private databases; as from April 1996, approximately 500 megabases (Mb) of nucleotide sequence were contained in GenBank, and approximately 200 000 sequences were held in the SWISS-PROT/Genpept/PIR database of non-redundant proteins. Combined with the wealth of biochemical information that is archived in public databases, it has become possible to describe rapidly the full repertoire of genes in a microbial genome, and to predict many of the metabolic pathways that an organism may utilize.

Progress in this field has been stimulated by the interests of the biotechnology and pharmaceutical industries in using genome-sequencing data as a basis for drug discovery. In turn, this has led to the development of proprietary databases containing genomic information, which provide the basis for *in silico* experiments to identify novel targets for drugs, and for laboratory experiments to identify genes that perform critical functions. This article summarizes some recent developments in this important area, focusing on bacterial sequences, and provides examples to illustrate how genomic-sequencing information from microbial pathogens can be used to select targets for vaccine and drug development. The overall process used to proceed from sequence generation to target validation is illustrated in Fig. 1.

D. R. Smith (smith@tnt.com) is at Genome Therapeutics Corporation, 100 Beaver Street, Waltham, MA 02154, USA.

Large-scale sequencing of bacterial genomes

Many laboratories use automated sample-preparation techniques and fluorescence-based gel readers [such as that produced by Applied Biosystems Inc., (ABI); Foster City, CA, USA] for the large-scale sequencing of bacterial genomes. These instruments have the advantage that they are efficient, and relatively easy to set up and operate. A few laboratories use computer-assisted multiplex sequencing to achieve the same end[1]. In multiplex sequencing, samples consisting of pools of up to 20 plasmids are processed through sample preparation and gel electrophoresis, and the resulting sequences are determined from electroblots of the gels by hybridization with radioactive or fluorescently labeled probes. This technique can be used to generate 40 films (or digitized images) from each sequencing gel. Although multiplex sequencing is efficient at producing large amounts of 'shotgun' data, it is more difficult to set up and operate in the laboratory than is fluorescence-based gel sequencing, and it is not suited to directed-finishing strategies. ABI machines are used in the author's laboratory to generate primer-directed reads for finishing and gap closure.

During the past year, a group at The Institute for Genomic Research (TIGR; Gaithersburg, MD, USA) reported the complete sequences of *Haemophilus*

 0167 – 7799/96/$15.00. PII: S0167-7799(96)10038-5

influenzae (1.8 Mb), a major cause of respiratory infections and meningitis, especially in children[2], and of *Mycoplasma genitalium* (0.6 Mb), which causes urethritis[3]. Approximately 1.6 Mb of contiguous sequence from the 4.7 Mb *Escherichia coli* genome has been published[4], and the sequencing of a further 2 Mb was reported at the 1995 Genome Sequencing and Analysis VII (GSA-VII) meeting[5]. The genome of *Helicobacter pylori* (1.7 Mb), the major cause of stomach ulcers, has been sequenced by Genome Therapeutics Corporation (GTC; Waltham, MA, USA) under a privately funded microbial-pathogen sequencing program. More than half (1.5 Mb) of the 2.8 Mb genome of *Mycobacterium leprae* (the etiologic agent of leprosy) has also been sequenced by GTC, and is available through GenBank, the GTC web site <http://www.tigr.com>, and through MycDB <http://www.biochem.kth.se/MycDB.html>, which contains mycobacterial genome mapping and sequence information[6].

Other microbial pathogens that are currently being sequenced include *Neisseria gonorrhoeae* (University of Oklahoma, Norman, OK, USA), *Streptococcus pyogenes* (University of Oklahoma), *Treponema pallidum* (University of Texas, Houston, TX, USA, and TIGR), *Mycobacterium tuberculosis* (GTC and the Sanger Centre, Hinxton, Cambridge, UK), and *Staphylococcus aureus* [GTC, and Human Genome Sciences (HGS; Rockville, MD, USA)].

In addition to these pathogens, the genomes of several archaebacteria and other non-pathogens are being sequenced. These include *Methanococcus janaschii* (TIGR), *Pyrococcus furiosis* (University of Utah, Salt Lake City, UT, USA), *Sulfolobus solfataricus* (Dalhousie University, Halifax, Nova Scotia, Canada), and *Pyrobaculum aerophilum* (California Institute of Technology, Pasadena, CA, USA, and University of California, Los Angeles, CA, USA). The 1.7 Mb genome of the archaeon *Methanobacterium thermoautotrophicum* is near completion at GTC (Ref. 7). Approximately 2 Mb of the 4.1 Mb *Bacillus subtilis* genome has now been sequenced by a consortium of European and Japanese laboratories, and the project may be completed by the end of 1996 (Ref. 8). Approximately 1 Mb of genomic sequence from the 2.7 Mb genome of the cyanobacterium *Synechocystis* sp. 6803 was recently published[9].

Within the next couple of years, therefore, we can expect an explosion of bacterial-genome sequence information from species representing a variety of phylogenetic lineages, including many pathogens.

Pharmaceutical companies have shown considerable interest in using pathogen genomics to facilitate the development of vaccines and small-molecule therapeutics. For example, researchers at GlaxoWellcome have sequenced a substantial fraction of the *H. pylori* genome to assist in the process of drug discovery. Over the past year, GTC has formed two research alliances with pharmaceutical companies to take advantage of sequences from microbial pathogens: one with Astra AB [illegible] new antibiotics and vaccines to treat *H. pylori* infection, and one with Schering-Plough (Union, NJ, USA), to develop broad-spectrum antibiotics and vaccines. Although the genomic route to drug discovery for bacterial pathogens is new and remains unproved, the basic paradigm (outlined below) of gene identification, followed by functional analysis and drug screening, is well established. Thus, it is likely that more companies will become involved, and that in the the future, additional research alliances between genomics companies and the pharmaceutical industry will materialize in this area.

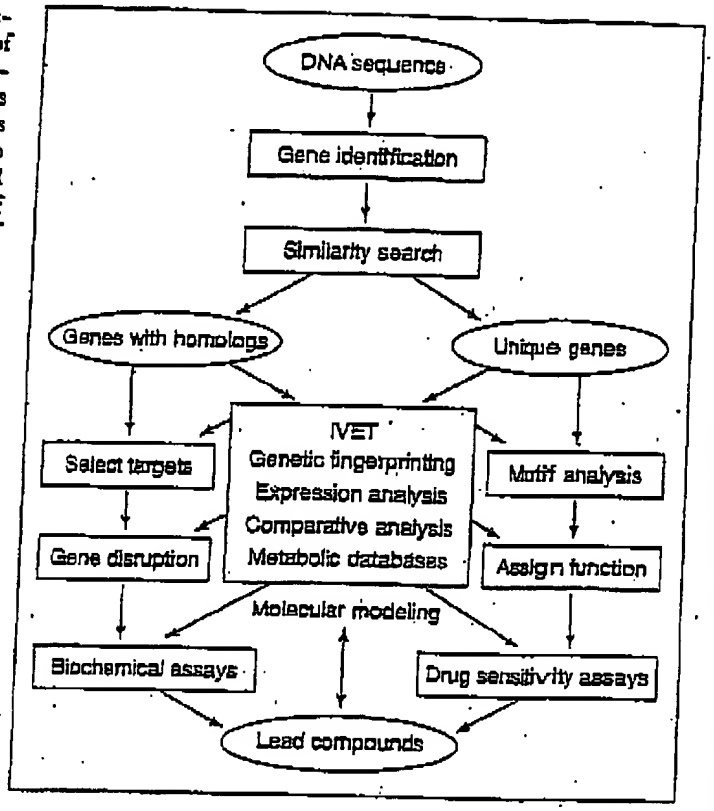
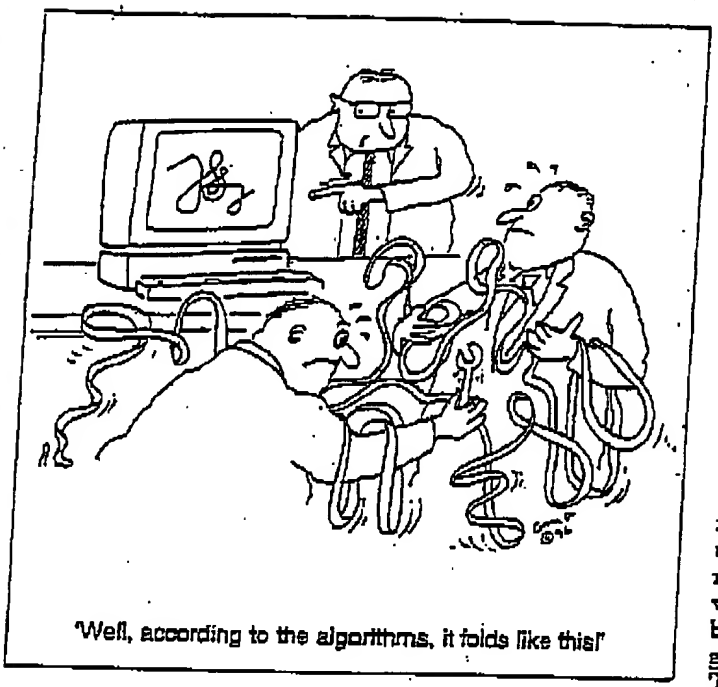


Figure 1

Flow diagram illustrating the process by which a microbial genome sequence is analysed and the information is used to direct experiments and aid in target selection for therapeutics development. The individual steps are referred to throughout the text. In the case of vaccine candidates, gene products from selected targets are expressed and tested in animal models.

From sequence to genes

The first task when confronted with an entire bacterial-genome sequence, is to identify all the genes. This can be accomplished using a variety of techniques, but the most successful approaches use a combination of reading-frame and codon-usage analysis, together with similarity searching, to identify putative genes with homology to previously described sequences. Commonly used tools include GeneMark[10], GenomeBrowser[11], BLAST (Ref. 12), and highly parallelized implementations of the Smith–Waterman



'Well, according to the algorithms, it folds like this!'

alignment, such as BLAZE, or MPsrch (Ref. 13). In general, organism-specific codon usage is highly predictive for bacterial genes, but its effective use depends on the existence of sufficient information to generate accurate codon-usage matrices. In some cases, subsets of genes within an organism will exhibit codon-usage patterns that deviate significantly from the norm[14]. Such genes are thought to represent evolutionarily recent acquisitions by phage transduction, conjugation, or some other form of horizontal transfer from other organisms. If enough of these genes are present, codon-usage tables of genomic subsets can be constructed to identify them. Translational start sites can be identified by the occurrence of start codons that coincide with abrupt changes in codon usage, the initiation of homology to previously characterized genes, or the presence of Shine–Dalgarno sequences[15]. Automated analysis tools (such as GenomeBrowser[11]) that provide a graphical display of open reading frames (ORFs), codon usage, database homologies and other features, make the task of identifying bacterial genes and their relationships with each other in the genome relatively straightforward. With the increasing pace of bacterial-genome sequencing, there is an emerging need for second-generation tools that will automate most of the laborious annotation process.

From genes to function

The second phase in the analysis of bacterial genomes is to identify the function of as many genes as possible. Currently, sequence homology is the most powerful tool. A high degree of homology between the putative translation product of a newly identified gene and an enzyme whose function has been thoroughly studied in other organisms, provides strong support for the function of that protein, especially if it is the only homolog in the genome under scrutiny. Other useful tools include programs that identify sequence motifs from databases such as PROSITE (Ref. 16), BLOCKS (Ref. 17), BEAUTY (Ref. 18) and ProDom (Ref. 19). If one is attempting to identify vaccine candidates, then examining highly expressed cell-surface proteins is relevant, so it is then useful to know whether a protein contains a secretion signal, even if nothing else is known about it. Although the tools described here are very good at identifying homologies, 25–40% of the genes in a bacterial genome typically fail to show significant similarity with known proteins.

Once the set of similarity-searching tools has been exhausted, one must return to molecular biology to further elucidate the function and expression pattern of predicted genes. Commonly used approaches to identifying essential genes in an organism include: the use of gene knockouts, disruptions using transposon-mediated mutagenesis, or homologous recombination with disrupted gene-constructs that contain an antibiotic-resistance cassette. Gene disruptions can be generated in a variety of ways, including sophisticated 'hit-and-run' approaches that interrupt a gene without introducing polar effects into downstream ORFs (Ref. 20). However, a gene-by-gene approach to the study of a whole genome is certainly time consuming and labor intensive.

The availability of large amounts of genome-sequence information has stimulated the development of new approaches to functional analysis on a genomic scale. This has been particularly true for researchers investigating yeast, where a concerted effort is being made to ascertain the function of every ORF in the genome. Such strategies include the conceptually simple, but technologically advanced, technique of making microarrays of polymerase chain reaction (PCR)-amplified gene sequences on glass slides to allow the fluorescence-based detection of quantitative hybridization signals from labeled cDNA probes on large numbers of genes simultaneously — perhaps even all the genes of an organism[21]. An ingenious PCR-based approach to efficient sequence-signature-based expression analysis has recently been demonstrated[22]. For example, a technique termed 'genetic fingerprinting' promises to replace individual gene knockouts by a global transposon-mutagenesis approach[23]. Insertions are induced *en masse* in a strain of interest, the strain is grown under a variety of conditions, and PCR products are analysed to identify genes in which transposon hops are under-represented because the genes are required for growth[23]. A conceptually similar dropout technique, which uses tagged transposons to identify the *Salmonella typhimurium* genes required for virulence in a mouse model, has been described[24].

Techniques that probe subsets of genes for a specific functionality, such as secretion or induction during growth in the host, have also been described. These techniques provide clones from which signature

sequences can be derived, so that corresponding genes can be identified by comparing them with the genomic sequence. The IVET (*in vivo* expression technology) technique, which detects gene fusions that result in the *in vivo* selectable expression of a defective *purA* gene or antibiotic-resistance marker, has been used to identify *Salmonella* genes, the expression of which is induced when the pathogen is grown in mice[25]. Finally, protein microsequencing[26] and mass-spectrometry-based peptide analysis[27] have been used to identify protein components (e.g. outer-membrane proteins) in partially purified mixtures, or to identify specific proteins separated by two-dimensional gel electrophoresis. Sequences generated in this manner can be used to correlate specific proteins with the gene sequences from which they are expressed.

Target selection and validation

The techniques described in the previous section can be used to identify genes in specific functional categories that may represent good targets for drug or vaccine development. In general, when developing new antibiotics, one is interested in genes that are essential under all growth conditions (and preferably even in quiescent cells), and for which inhibitors with useful chemical properties, such as permeability and low toxicity, can be identified. One advantage of having the entire sequence of a genome is that targets can be prioritized in terms of their activities and the properties of compounds that are known to interact with them. Even with the results of knockout or *in vivo* expression experiments, additional biological information can aid in narrowing down the field of choices. For example, genes can be selected on the basis of their probable roles in intracellular metabolism. Databases, such as EcoCyc (Ref. 28) or PUMA (Ref. 29), that describe known metabolic pathways can be helpful in this regard. Detailed structural information about homologs of identified genes (determined using the Protein DataBank[30]) can be used to assist in the molecular modeling of inhibitors (some resources for molecular modeling can be found at Ref. 31).

As more genomes are sequenced, it will become possible to identify genes that are unique to a particular organism or group of organisms, or genes that are conserved in certain groups. Thus, for example, it will be possible to use electronic comparison to identify genes that are present in *H. pylori* but not in other gut-dwelling bacteria such as *E. coli*, providing a basis for the development of antibiotics specific to *H. pylori*. Although combinatorial chemistries promise to speed up our ability to synthesize and screen large numbers of unique chemical entities, the sequence-based approach described here provides an avenue for the rational identification and selection of key targets for therapeutics development. Ultimate validation of the targets will, of course, require additional experiments such as protein expression, biochemical-assay development and animal studies to identify those with the most useful properties or inhibitors.

Acknowledgements

The sequencing of *Mycobacterium leprae* and *M. tuberculosis*, and technology development for multiplex sequencing is supported by a NIH Genome Science and Technology Center grant 1P01-HG1106-01 from the National Center for Human Genome Research. The sequencing of *Methanobacterium thermoautotrophicum* is supported under the Microbial Genome Program by Grant No. DE-FC02-95ER.61967 from the Office of Health and Environmental Research of the US Department of Energy. The sequencing of *Helicobacter pylori* and *Staphylococcus aureus* is supported by Genome Therapeutics Corporation. Thanks to Brad Gould for comments on the manuscript.

References

1 Church, G. M. and Kieffer-Higgins, S. (1988) *Science* 240, 185–188
2 Fleischmann, R. D. *et al.* (1995) *Science* 269, 496–512
3 Fraser, C. D. *et al.* (1995) *Science* 270, 397–403
4 Burland, V., Plunkett, G., III, Sofia, H. J., Daniels, D. L. and Blattner, F. R. (1995) *Nucleic Acids Res.* 23, 2105–2119
5 Burland, V., Plunkett, G., III and Blattner, F. R. (1995) in *Genome Science and Technology* 1, P–16, Mary Ann Liebert
6 Bergh, S. and Cole, S. T. (1994) *Mol. Microbiol.* 12, 517–534
7 Smith, D. R. *et al.* (1995) in *Genome Science and Technology* 1, P–48, Mary Ann Liebert
8 Devine, K. (1995) *Trends Biotechnol.* 13, 210–216
9 Kaneko, T. *et al.* (1995) *DNA Res.* 2, 153–166
10 Borodovsky, M. and Mclninch, J. (1993) *Comput. Chem.* 17, 123–133
11 Robison, K. R. and Church, G. M. (1995) <http://www.helmont.com/gb.html>
12 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410
13 MPath <http://www.ebi.ac.uk/mathes/blitz.html>
14 Medigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991) *J. Mol. Biol.* 222, 851–856
15 Shine, J. and Dalgarno, L. (1975) *Eur. J. Biochem.* 57, 221–230
16 Bairoch, A. (1991) *Nucleic Acids Res.* 19, 2241–2245
17 Henikoff, S. and Henikoff, J. G. (1991) *Nucleic Acids Res.* 19, 6565–6572
18 Wooley, K. C., Wiese, B. A. and Smith, R. F. (1995) *Genome Res.* 5, 173–184
19 Sonnhammer, E. L. and Kahn, D. (1994) *Protein Sci.* 3, 482–492
20 Link, A. J. and Church, G. M. <http://sweed.med.harvard.edu/12hgs/pKOS.html>
21 Schena, M., Shalon, D., Davis, R. D. and Brown, P. O. (1995) *Science* 270, 467–470
22 Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) *Science* 270, 484–487
23 Smith, V., Botstein, D. and Brown, P. O. (1995) *Proc. Natl Acad. Sci. USA* 92, 5479–6483
24 Hensel, M. *et al.* (1995) *Science* 269, 400–403
25 Mahan, M. J. *et al.* (1995) *Proc. Natl Acad. Sci. USA* 92, 669–673
26 Tempst, P., Link, A. J., Riviere, L. R., Fleming, M. and Elicone, C. (1990) *Electrophoresis* 11, 537–553
27 James, P., Quadroni, M., Carafoli, E. and Gonnet, G. (1993) *Biochem. Biophys. Res. Commun.* 195, 58–64
28 Karp, P. D. (1992) *CABIOS* 8, 347–357
29 Grantcharoff, T., Maltsev, N., Overbeek, R., Selkov, E. <http://www.mcs.anl.gov/home/compbio/PUMA>
30 Protein DataBank <http://www.pdb.bnl.gov>
31 <http://www.pharmacy.wisc.edu>